smfd2(13a).tex

Day 2. 17.11.2013

3. Large-sample properties of maximum-likelihood estimators

We assume the following regularity conditions:
(i) differentiability under the integral sign twice (as before);
(ii) finite positive Fisher information per reading $I(\theta)$;
(iii) In some neighbourhood N of the true parameter value θ_0,

$$\Big|\frac{\partial^3}{\partial\theta^3}\log f(\mathbf{x};\theta)\Big| \le H(\mathbf{x}), \quad \text{where} \quad \sup_{\theta\in N} E_\theta H(\mathbf{X}) \le M < \infty.$$

Theorem (Cramér, 1946). Under the above regularity conditions, the MLE $\hat\theta$ of the true parameter value θ_0 is
(i) *strongly consistent*: $\hat\theta \to \theta_0$ as $n \to \infty$, a.s.,
(ii) *asymptotically efficient*: *var* $\hat\theta \sim 1/(nI(\theta))$, the Cramér-Rao lower bound;
(iii) *asymptotically normal*: $\sqrt{nI(\theta_0)}(\hat\theta - \theta_0) \to \Phi = N(0,1)$ $(n \to \infty)$.

Proof. (i) We use Taylor's theorem to expand the score function $s = \ell'$ about $\theta = \theta_0$:

$$s(\theta) = s(\theta_0) + (\theta-\theta_0)s'(\theta_0) + \frac{1}{2}(\theta-\theta_0)^2 s''(\theta^*),$$

for some θ^* between θ_0 and θ. Since $\ell(x_1,\dots,x_n;\theta) = \sum_1^n \ell(x_i;\theta)$, and similarly for $s = \ell'$, this says (on dividing by n)

$$s(\mathbf{x};\theta)/n = \frac{1}{n}\sum_1^n s(x_i;\theta) = \frac{1}{n}\sum_1^n s(x_i;\theta_0) + (\theta-\theta_0).\frac{1}{n}\sum_1^n s'(x_i;\theta_0)$$

$$+\frac{1}{2}(\theta-\theta_0)^2.\frac{1}{n}\sum_1^n s''(x_i;\theta^*). \qquad (*)$$

The first term on the RHS is an average of iid rvs with mean $Es(\theta_0) = 0$, by (a) of I.2. So by SLLN, this $\to 0$ a.s. as $n \to \infty$. Similarly, by SLLN the second term on RHS converges a.s. to

$$(\theta-\theta_0)s'(\theta_0) = -I(\theta_0)(\theta-\theta_0).$$

The third term on RHS of $(*)$ is bounded by $\frac{1}{2}M(\theta-\theta_0)^2$, by our regularity assumption (iii) (as $s'' = \ell'''$). For θ close enough to θ_0, this is negligible wrt the second term. So RHS $\sim$ second term:

$$\text{RHS} \sim -I(\theta_0).(\theta-\theta_0).$$

Since $I(\theta_0) \in (0, \infty)$, by (ii), the *sign* of the RHS is thus *opposite* to that of $\theta - \theta_0$, for large enough n and θ close enough to θ_0. For such n and θ, RHS *changes sign* in every neighbourhood of θ_0 (just take θ through θ_0). But LHS = RHS, so the LHS too changes sign in every neighbourhood of θ_0, for large enough n. This says that there is a root $\hat{\theta} = \hat{\theta_n}$ of the *likelihood equation*

$$s(\mathbf{x}; \theta) = 0 \qquad (LE)$$

in every neighbourhood of θ_0.

Since this neighbourhood can be arbitrarily small, we must have

$$\hat{\theta} = \hat{\theta_n} \to \theta_0 \qquad (n \to \infty) \qquad a.s.,$$

proving the strong consistency of the MLE $\hat{\theta}$ and (i).

Now put $\theta = \hat{\theta}$ in $(*)$. The LHS is 0, by definition of MLE (recall $s = \ell' = (\log L)'$). The third term on RHS is negligible wrt the second term, because of the extra factor $\theta - \theta_0 \to 0$, by (i). So we can neglect this term, leaving

$$0 \sim \frac{1}{n} \sum_1^n s(x_i; \theta_0) + (\hat{\theta} - \theta_0) . \frac{1}{n} \sum_1^n s'(x_i; \theta_0).$$

Rearranging,

$$\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0) \sim \frac{[\sum_1^n s(x_i; \theta_0)] / \sqrt{nI(\theta_0)}}{[\sum_1^n -s'(x_i; \theta_0)]/(nI(\theta_0))} \qquad (n \to \infty). \qquad (**)$$

By CLT, the numerator on RHS $\to \Phi = N(0, 1)$, as $s(x_i; \theta_0)$ are iid with mean 0 and variance $I(\theta_0)$. By LLN, the denominator on RHS $\to 1$ a.s., as $-s'(x_i; \theta_0)$ are iid with mean $I(\theta_0)$ (by I.2, Prop.). Combining, RHS $\to \Phi$. So LHS $\to \Phi$:

$$\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0) \to \Phi = N(0, 1) : \qquad var\ (\hat{\theta} - \theta_0) = var\ \hat{\theta} \sim 1/(nI(\theta_0)),$$

the Cramér-Rao bound, so $\hat{\theta}$ is asymptotically efficient, proving (ii), *and* $\hat{\theta}$ is asymptotically normal, proving (iii). //

The regularity conditions above can be weakened, at the cost of a harder proof, but *some* regularity conditions are needed. The phrase "under suitable regularity conditions" recurs with remorseless regularity in textbook treatments of large-sample properties of MLEs.

Example. Uniform $U(a,b)$, or $U(\theta - \frac{1}{2}, \theta + \frac{1}{2})$. See IS II, where we showed that here the MLEs converge at a different rate (n, not $\sqrt{n}$) and to a different limit (exponential, or symmetric exponential, not normal).

Vector Parameters.

If $\theta = (\theta_1, \ldots, \theta_r)$ is an r-dimensional (vector) parameter, one can proceed as above. We now obtain the (Fisher) *information matrix*

$$I(\theta) = (I_{ij}(\theta))_{i,j=1}^{r},$$

where we use suffix notation for partial differentiation ($g_i := \partial g / \partial \theta_i$, etc.) and

$$I_{ij}(\theta) := E[\ell_i(\theta)\ell_j(\theta)] = E[-\ell_{ij}(\theta)].$$

Under regularity conditions as above (we assume the information matrix is positive definite, so we can invert it), we again obtain consistency, asymptotic efficiency and asymptotic normality:

$$\hat{\theta} \sim N_r(\theta, n^{-1} I^{-1}(\theta)).$$

Stochastic process versions.

The true context for results such as the above is not random variables as above, but stochastic processes. Infinite-dimensional versions are possible (and needed), in which conclusions are drawn, not for one time-point at a time, but for infinitely many together – say, all $t \in [0,1]$, or all $t \geq 0$. We shall develop these ideas later (Day 3). Meanwhile, we mention our main source for such things,

[vdVW] Aad van der VAART and Jon A. WELLNER, *Weak convergence and empirical processes, with applications to statistics*, Springer, 1996.

In particular, [vdVW] contains detailed accounts of *M-estimators* (3.2) ('M for maximum' – generalising MLEs), and *Z-estimators* (3.3) ('Z for zero': the MLE is a zero of the *likelihood equation* $\ell' = 0$).

Fisher's method of scoring.

Here we replace $s'(\theta^*)$ by $E[s'(t)] = -I(t)$ (we know $\theta^* \sim t \sim \theta_0$ by above). Then our next (better) approximation is

$$\hat{\theta} \sim t - s(t)/E[s'(t)] = t + s(t)/I(t).$$

This is Fisher's *method of scoring.* As always with iterations: to implement this numerically, one needs a "do-loop" (while .... do ...., else stop).

Exercise. Implement this in C++ (the "official programming language" for this course), for the Cauchy location family (Problems 3 Q3). First, choose a μ (arbitrarily – or, by sampling from a chosen distribution). Then, sample from the Cauchy distribution with this μ. Then, perform the above iterations (by either, or better still both, of the Newton-Raphson and Fisher methods), to estimate this μ from the data.

Reparametrisation and the Delta Method.

Suppose we are using parameter θ, but wish to change to some alternative parametrisation, $g(\theta)$, where g is continuously differentiable. A CLT for θ such as

$$\sqrt{n}(T_n - \theta) \to N(0, \sigma(\theta)^2)$$

(as holds above, with T_n the MLE $\hat{\theta}$ based on a sample of size n and $\sigma^2(\theta) = 1/I(\theta)$) transforms into a CLT for $g(\theta)$:

$$\sqrt{n}(g(T_n) - g(\theta)) \to N(0, [g'(\theta)\sigma(\theta)]^2).$$

For,

$$g(T_n) - g(\theta) = (T_n - \theta)(g'(\theta) + \epsilon_n),$$

with ϵ_n a (random) error term. One can show this to be negligible for large n, so

$$g(T_n) - g(\theta) \sim (T_n - \theta)g'(\theta).$$

Since $var(cX) = c^2 \ var \ X$, the result follows.

This is called the *delta method*, and is often useful. It can be extended from random variables to stochastic processes (i.e., from one or finitely many to infinitely many dimensions), and we shall meet it again later.

4. Sufficiency and Minimal Sufficiency

Recall (IS II) the idea of sufficiency as data reduction, and minimal sufficiency as data reduction carried as far as possible without loss of information. We now formalise this.

Definition (Fisher, 1922). To estimate a parameter θ from data $\mathbf{x}$, a statistic $T = T(\mathbf{x})$ is *sufficient* for θ if the conditional distribution of $\mathbf{x}$ given $T = T(\mathbf{x})$ does not depend on θ.

Interpretation. Always use what you know. We know T: is this enough? The conditional distribution of $\mathbf{x}$ given T represents the information remaining in the data $\mathbf{x}$ over and above what is in the statistic T. If this does not involve θ, the data *cannot* have anything left in it to tell us about θ beyond what is already in T.

The usual – because the easiest – way to tell when one has a sufficient statistics is the result below. The sufficiency part is due to Fisher in 1922, the necessity part to J. NEYMAN (1894-1981) in 1925.

Theorem (Factorisation Criterion; Fisher-Neyman Theorem. T is sufficient for θ if the likelihood factorises:

$$f(\mathbf{x};\theta) = g(T(\mathbf{x});\theta)h(\mathbf{x}),$$

where g involves the data only through T and h does not involve the parameter θ.

Proof. We give the discrete case; the density case is similar.
Necessity. If such a factorisation exists,

$$P_\theta(\mathbf{X}=\mathbf{x}) = g(T(\mathbf{x}),\theta)h(\mathbf{x}),$$

then given t_0,

$$P(T=t_0) = \sum_{\mathbf{x}:T(\mathbf{x})=t_0} P_\theta(\mathbf{X}=\mathbf{x}) = \sum_{\mathbf{x}:T(\mathbf{x})=t_0} g(T(\mathbf{x}),\theta)h(\mathbf{x}) = g(t_0,\theta)\sum_{\mathbf{x}:T(\mathbf{x})=t_0} h(\mathbf{x}).$$

So $P_\theta(\mathbf{X}=\mathbf{x}|T=t_0) = P_\theta(\mathbf{X}=\mathbf{x}\ \&\ T=T(\mathbf{X})=t_0)/P_\theta(T=t_0)$ is 0 unless $T(\mathbf{x})=t_0$, in which case it is

$$P_\theta(\mathbf{X}=\mathbf{x})/P_\theta(T=t_0) = \frac{g(t_0;\theta)h(\mathbf{x})}{g(t_0;\theta)\sum_{T(\mathbf{x})=t_0} h(\mathbf{x})} = \frac{h(\mathbf{x})}{\sum_{T(\mathbf{x})=t_0} h(\mathbf{x})}.$$

This is independent of θ, so T is sufficient.
Sufficiency. If T is sufficient, the conditional distribution of $\mathbf{X}$ given T is independent of θ:

$$P_\theta(\mathbf{X}=\mathbf{x}|T=t_0) = c(\mathbf{x},t_0), \qquad \text{say.} \qquad (i)$$

The LHS is $P(\mathbf{X}=\mathbf{x}\ \&\ T(\mathbf{X})=t_0)/P(T=t_0)$. Now the numerator is 0 unless $t_0=T(\mathbf{X})$. Defining $c(\mathbf{x},t_0)$ to be 0 unless $t_0=T(\mathbf{x})$, we have (i) in all cases, and now

$$c(\mathbf{x},t_0) = P_\theta(\mathbf{X}=\mathbf{x})/P(T(\mathbf{X})=t_0),$$

as "& $T(\mathbf{X}) = t_0 = T(\mathbf{x})$" is redundant. So now

$$P_\theta(\mathbf{X} = \mathbf{x}) = P_\theta(T(\mathbf{X}) = t_0)c(\mathbf{x}, t_0),$$

a factorisation of the required type. //

Cor. If $U = a(T)$ with a injective (one-to-one), T sufficient implies U sufficient.

Proof. $T = a^{-1}(U)$ as a is one-to-one, so

$$f(\mathbf{x}; \theta) = g(a^{-1}(U); \theta)h(\mathbf{x}) = G(U(\mathbf{x}); \theta)h(\mathbf{x}),$$

say, a factorisation of Fisher-Neyman type, so U is sufficient. //

So if, e.g. T is sufficient for the population variance σ^2, $\sqrt{T}$ is sufficient for the standard deviation σ, etc.

Example: Normal families $N(\mu, \sigma^2)$.
(i) The joint likelihood factorises into the product of the marginal likelihoods, so

$$f(\mathbf{x}; \mu, \sigma^2) = \frac{1}{(2\pi)^{\frac{1}{2}n}\sigma^n}. \exp\{-\frac{1}{2}\sum_1^n (x_i - \mu)^2/\sigma^2\}. \qquad (1)$$

Since $\bar{x} := \frac{1}{n}\sum_1^n x_i$, $\sum(x_i - \bar{x}) = 0$, so

$$\sum(x_i-\mu)^2 = \sum[(x_i-\bar{x})+(\bar{x}-\mu)]^2 = \sum(x_i-\bar{x})^2+n(\bar{x}-\mu)^2 = n(S^2+(\bar{x}-\mu)^2) :$$

the likelihood is

$$L = f(\mathbf{x}; \mu, \sigma^2) = \frac{1}{(2\pi)^{\frac{1}{2}n}\sigma^n}. \exp\{-\frac{1}{2}n(S^2 + (\bar{x} - \mu)^2)/\sigma^2\}. \qquad (2)$$

By the Factorisation Criterion, $(\bar{x}, S^2)$ is (jointly) sufficient for (μ, σ^2). So for a *normal* family: only *two* numbers are needed for the two parameters μ, σ^2, namely $\bar{x}, S^2$ (equivalently, $\sum X, \sum X^2$ – note that good programmable pocket calculators have keys for $\sum X, \sum X^2$ for this purpose!)
(ii) Now suppose σ is *known* (so counts as a constant, not a parameter). Then (2) says that *$\bar{x}$ is now sufficient for μ.*
(iii) Now suppose μ is known. Then (1) says that now $\sum(x_i - \mu)^2$ *is sufficient*

for σ^2.

Minimal Sufficiency. Sufficiency enables *data reduction* – reducing from n numbers (n is the sample size – the bigger the better) to a much smaller number (as above). Ideally, we would like to reduce as much as possible, without loss of information. How do we know when we have done this?

Recall that when applying a function, we lose information in general (we do not lose information only when the function is injective – one-to-one, when we can go back by applying the inverse function). This leads to the following

Definition. A sufficient statistic T is *minimal (sufficient)* for θ if T is a function of any other sufficient statistic T'.

Minimal sufficient statistics are clearly desirable ('all the information with no redundancy'). The following result gives a way of constructing them.

Theorem (Lehmann & Scheffé, 1950). If T is such that the likelihood ratio $f(\mathbf{x};\theta)/f(\mathbf{y};\theta)$ is independent of θ iff $T(\mathbf{x}) = T(\mathbf{y})$, then T is a minimal sufficient statistic for θ.

We quote this. To find minimal sufficient statistics, we form the likelihood ratio, and seek to eliminate the parameters. This works very well in practice, as examples show (see Problems 2).

5. Location and scale; Tails

In one dimension, the mean μ gives us a natural measure of *location* for a distribution. The variance σ^2, or standard deviation (SD) σ, give us a natural measure of *scale.*

Note. The variance has much better mathematical properties (e.g., it adds over independent, or even uncorrelated, summands). But the SD has the *dimensions* of the random variable, which is better from a physical point of view. As moving between them is mathematically trivial, we do so at will, without further comment.

Example: Temperature. In the UK, before entry to the EU (or Common Market as it was then), temperature was measured in degrees Fahrenheit, F (freezing point of water 32^oF, boiling point 212^oF (these odd choices are only of historical interest – but dividing the freezing-boiling range into 180 parts rather than 100 is better attuned to homo sapiens being warm-blooded, and most people having trouble with decimals and fractions!) The natural choice for freezing is 0; 100 parts for the freezing-boiling range is also natural when using the metric system – whence the Centigrade (= Celsius) scale. Back then, one used F for ordinary life, C for science, and the conversion rules

$$C = \frac{5}{9}(F - 32), \qquad F = \frac{9}{5}C + 32$$

were part of the lives of all schoolchildren (and the mechanism by which many of them grasped the four operations of arithmetic!)
Pivotal quantities.

A *pivotal quantity*, or *pivot*, is one whose distribution is independent of parameters. Pivots are very useful in forming *confidence intervals.*

Defn. A *location family* is one where, for some reference density f, the density has the form $f(x - \mu)$; here μ is a *location parameter.* A *scale family* (usually for $x \geq 0$) is of the form $f(x/\sigma)$; here σ is a *scale parameter.* A *location-scale family* is of the form $f(\frac{x-\mu)}{\sigma})$.
Pivots here are

$$\bar{X} - \mu \quad \text{(location)}; \quad \bar{X}/\sigma \quad \text{(scale)}; \quad \frac{\bar{X} - \mu}{\sigma} \quad \text{(location-scale)}.$$

Examples. The normal family $N(\mu, \sigma^2)$ is a location-scale family.
The *Cauchy location family* is

$$f(x - \mu) = \frac{1}{\pi[1 + (x - \mu)^2]}.$$

In higher dimensions, the location parameter is the mean μ (now a *vector*); the scale parameter is now the *covariance matrix*

$$\Sigma = (\sigma_{ij}), \qquad \sigma_{ij} := cov(X_i, X_j) = E[(X_i - EX_i)(X_j - EX_j)].$$